Abstract

A system and method for use with a data mining application for a large database having a large number of records. A selection attribute is chosen from one of a plurality of attributes contained by records within the database. Records are scanned in the database and a randomizing function is applied to the selection attribute of each record to create a randomized record value. A selection criteria is then applied to identify records for inclusion within a subset of records (smaller than the original data set) by comparing the randomized record value of each record with the selection criteria. The subset of records having a randomized record value satisfying the selection criteria approximates the entire database but takes up less memory and can be evaluated or scanned much more quickly.